(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 April 2002 (11.04.2002)

PCT

(10) International Publication Number
WO 02/29780 A2

(51) International Patent Classification[7]: G10L

(21) International Application Number: PCT/US01/31121

(22) International Filing Date: 3 October 2001 (03.10.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/238,560 4 October 2000 (04.10.2000) US

(71) Applicant *(for all designated States except US)*: **CLARITY, LLC** [US/US]; 3290 West Big Beaver Road, Suite 200, Troy, MI 48084 (US).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **ERTEN, Gamze** [US/US]; 1848 Elk Lane, Okemos, MI 48864 (US).

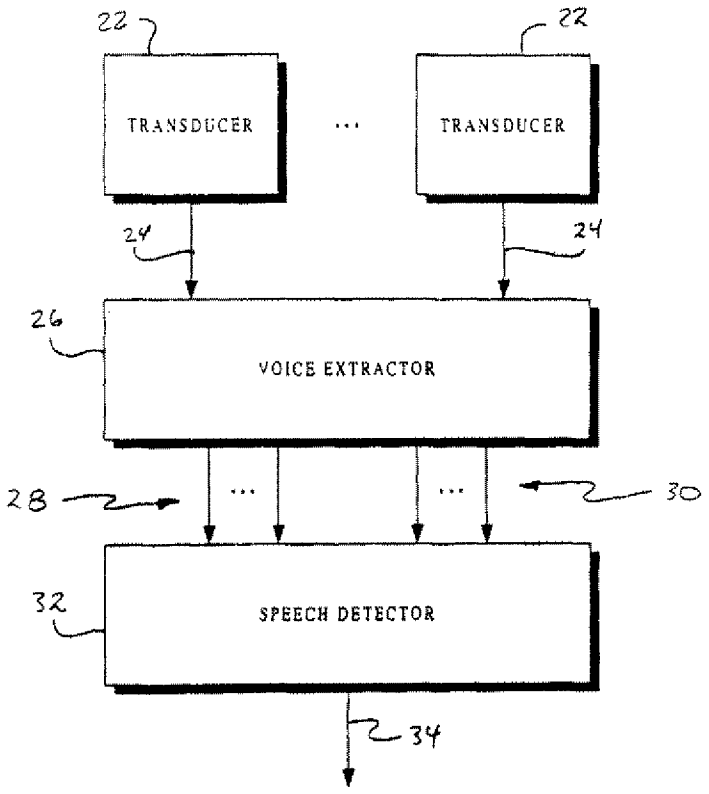
(74) Agents: **CHUEY, Mark, D.** et al.; Brooks & Kushman, 1000 Town Center, Twenty-Second Floor, Southfield, MI 48075 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF,

[Continued on next page]

(54) Title: SPEECH DETECTION



(57) Abstract: Speech in the presence of noise is detected by first extracting at least one extracted speech signal (28) from at least one received signal (24) and extracting at least one extracted noise signal (30) from the at least one received signal (24). A detected speech signal (34) is generated based on both at least one extracted speech signal (28) and on at least one extracted noise signal (30).

WO 02/29780 A2

CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

SPEECH DETECTION

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to detecting the presence of speech.

2. Background Art

Speech detection is the process of determining whether or not a certain segment of recorded or streaming audio signal contains a voice signal. The voice signal typically is a voice signal of interest which may appear in the presence of noise including other voice signals. Speech detection may be used in a wide variety of applications including speech activated command and control systems, voice recording, voice coding, voice transmitting systems such as telephones, and the like.

A barrier to the proliferation and user acceptance of voice based command and communications technologies has been noise sources that contaminate the speech signal and degrade the quality of speech processing results. The consequences are poor voice signal quality, especially for far field microphones, and low speech recognition accuracy for voice based command applications. The current commercial remedies, such as noise cancellation filters and noise cancelling microphones, have been inadequate to deal with a multitude of real world situations.

Elimination of noise from an audio signal leads to better speech detection. If noise mixed into the signal is reduced, while eliminating little or none of the voice component of the signal, a more straight forward conclusion as to whether a certain part of the signal contains voice may be made.

Speech detection can be based on several criteria. One commonly used criteria is the power of the signal. This approach assumes that the speaker is within a short distance from the microphone so that when the speaker speaks, the power of the signal recorded by the transducer that senses or registers the sound will rise significantly. These methods take advantage of the fact that speech is intermittent. Due to this intermittence, as well as the proximity of the speaker to the microphone, gaps between utterances will contain lower levels of signal power then the proportions that contain speech. A problem with such techniques is that speech itself does not generate a constant power. Thus, the surge in power of the signal will be less for speech that is not voiced. Speech detection based on signal power works best when the noise level is significantly lower then the speech level. However, such techniques tend to fail in the presence of medium or high levels of noise.

SUMMARY OF THE INVENTION

Speech detection of the present invention relies on characteristics of the estimated speech and on characteristics of estimated noise. Speech detection is based on speech signals and noise signals which are at least partially separated from each other.

A speech detection system is provided. The system includes at least one transducer converting sound into an electrical signal. A voice extractor produces at least one extracted speech signal and at least one extracted noise signal based on the electrical sound signals. A speech detector generates a detected speech signal based on the at least one extracted speech signal and on the at least one extracted noise signal. The speech detector may recognize periods of speech based on at least one property of the extracted speech signal and on at least one corresponding property of the at least one extracted noise signal.

Periods of speech may be recognized based on statistical properties, spectral properties, estimated relative proximity of a speaker to at least two of the transducers, an envelope of the extracted speech signal, signal power, and the like.

In an embodiment of the present invention, the at least one extracted speech signal is divided in time into a plurality of windows. The speech detector generates the detected speech signal based on determining whether or not speech is present in each window. The at least one extracted speech signal may be divided into a plurality of frequency bands with the speech detector determining whether or not speech is present in each frequency band for each window. The detected speech signal may then be based on a combination of the determination for each frequency band for each window.

In another embodiment of the present invention, a variable rate coder changes coding rate for coding the detected speech signal based on a determined presence of speech in the detected speech signal.

In still another embodiment of the present invention, a variable rate compressor changes compression rate for compressing the detected speech signal based on a determined presence of speech in the detected speech signal.

A method of detecting speech in the presence of noise is also provided. At least one signal containing speech mixed with noise is received. At least one extracted speech signal is extracted from the received signal. At least one extracted noise signal is also extracted from the received signal. A detected speech signal is generated based on at least one extracted speech signal and on at least one extracted noise signal.

In an embodiment of the present invention, the detected speech signal includes periods where the extracted speech signal is attenuated.

In another embodiment of the present invention, the detected speech signal includes a likelihood of speech presence.

A method of detecting speech is also provided. At least one noise signal is received. At least one speech signal having a greater content of speech then the at least one noise signal is also received. At least one noise parameter is

extracted from the noise signal. At least one speech parameter is extracted from the speech signal. The at least one speech parameter and the at least one noise parameter are compared and the presence of speech is detected based on this comparison.

Another method of detecting speech is provided. A noise signal and a speech signal having a greater speech content then the noise signal are received. The speech signal is divided into a plurality of speech frequency bands. The noise signal is divided into a plurality of noise frequency bands, each noise frequency band corresponding to one of the speech frequency bands. For each speech frequency band, at least one detection parameter is calculated based on at least one property of the speech frequency band and on at least one property of the corresponding noise frequency band. A frequency band output is generated based on the at least one detection parameter.

The above objects and other objects, features, and advantages of the present invention are readily apparent from the following detailed description of the best mode for carrying out the invention when taken in connection with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIGURE 1 is a block diagram of a speech detection system according to an embodiment of the present invention;

FIGURE 2 is a block diagram of signal separation according to an embodiment of the present invention;

FIGURE 3 is a block diagram of a feed-forward state space architecture for signal separation according to an embodiment of the present invention;

FIGURE 4 is a block diagram of a feed-back state space architecture for signal separation according to an embodiment of the present invention;

FIGURE 5 is a block diagram of a two transducer voice extractor having a plurality of extracted speech signal outputs according to an embodiment of the present invention;

FIGURE 6 is a block diagram of a two transducer voice extractor generating one extracted speech signal and one extracted noise signal according to an embodiment of the present invention;

FIGURE 7 is a block diagram illustrating a voice detector according to an embodiment of the present invention;

FIGURE 8 is a block diagram illustrating a voice detector using multiple frequency bands according to an embodiment of the present invention;

FIGURE 9 is a histogram plot of a typical voice signal;

FIGURE 10 is a histogram plot of typical noise signal;

FIGURE 11 is a frequency plot of a typical voice signal;

FIGURE 12 is a frequency plot of a typical noise signal;

FIGURE 13 is schematic diagram illustrating relative transducer placement for proximity-based speech detection according to an embodiment of the present invention;

FIGURE 14 is a plot of a noisy speech signal;

FIGURE 15 is a plot of a speech detective signal according to an embodiment of the present invention; and

FIGURE 16 is a block diagram illustrating compressing or coding according to an embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT(S)

Referring to Figure 1, a block diagram illustrating a speech detection system according to an embodiment of the present invention is shown. A speech detection system, shown generally by 20, includes one or more transducers 22 converting sound into sound signals 24. Typically, transducers 22 are microphones and sound signals 24 are electrical signals. Voice extractor 26 receives sound signals 24 and generates at least one extracted speech signal 28 and at least one extracted noise signal 30. Extracted speech signals 28 contain a greater content of desired speech then do extracted noise signals 30. Likewise, extracted noise signals 30 contain a greater noise content then do extracted speech signals 28. Thus, extracted speech signals 28 are "speechier" than extracted noise signals 30 and extracted noise signals 30 are "noisier" than extracted speech signals 28. Speech detector 32 receives at least one extracted speech signal 28 and at least one extracted noise signal 30. Speech detector 32 generates detected speech signal 34 based on received extracted speech signals 28 and on extracted noise signals 30.

Detected speech signal 34 may take on a variety of forms. For example, detected speech signal 34 may include one or more extracted speech signals 28, or combinations of extracted speech signals 28, in which periods where speech has not been detected are attenuated. Detected speech signal 34 may also include one or more signals indicating a likelihood of speech presence in one or more extracted speech signals 28 or sound signals 24.

Referring now to Figure 2, a block diagram of signal separation according to an embodiment of the present invention is shown. Signal separation permits one or more signals, received by one or more sound sensors, to be separated from other signals. Signal sources 40 indicated by $\mathbf{s}(t)$, represents a collection of source signals, including at least one desired voice signal, which are intermixed by mixing environment 42 to produce mixed signals 44, indicated by $\mathbf{m}(t)$. Voice

extractor 26 extracts one or more extracted speech signals 28 and one or more extracted noise signals 30 from mixed signals 44 to produce a vector of separated signals 46 indicated by $\mathbf{y}(t)$.

Many techniques are available for signal separation. One set of techniques is based on neurally inspired adaptive architectures and algorithms. These methods adjust multiplicative coefficients within voice extractor 26 to meet some convergence criteria. Conventional signal processing approaches to signal separation may also be used. Such signal separation methods employ computations that involve mostly discrete signal transforms and filter/transform function inversion. Statistical properties of signals 40 in the form of a set of cumulants are used to achieve separation of mixed signals where these cumulants are mathematically forced to approach zero. Additional techniques for signal separation are described in U.S. Patent Applications 09/445,778 filed March 10, 2000; 09/701,920 filed December 4, 2000; and 09/823,586 filed March 30, 2001; and PCT publications WO 98/58450 published December 23, 1998 and WO 99/66638 published December 23, 1999; each of which is herein incorporated by reference in its entirety.

Mixing environment 42 may be mathematically described as follows:

$$\dot{\bar{\mathbf{X}}} = \bar{\mathbf{A}}\,\bar{\mathbf{X}} + \bar{\mathbf{B}}\,\mathbf{s}$$
$$\mathbf{m} = \bar{\mathbf{C}}\,\bar{\mathbf{X}} + \bar{\mathbf{D}}\,\mathbf{s}$$

where $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$, $\bar{\mathbf{C}}$ and $\bar{\mathbf{D}}$ are parameter matrices and $\bar{\mathbf{X}}$ represents continuous-time dynamics or discrete-time states. Voice extractor 26 may then implement the following equations:

$$\dot{\mathbf{X}} = \mathbf{A}\,\mathbf{X} + \mathbf{B}\,\mathbf{m}$$
$$\mathbf{y} = \mathbf{C}\,\mathbf{X} + \mathbf{D}\,\mathbf{m}$$

where $\mathbf{y}$ is the output, $\mathbf{X}$ is the internal state of voice extractor 26, and $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ are parameter matrices.

Referring now to Figures 3 and 4, block diagrams illustrating state space architectures for signal mixing and signal separation are shown. Figure 3 illustrates a feedforward voice extractor architecture 26. Figure 4 illustrates a feedback voice extractor architecture 26. The feedback architecture leads to less restrictive conditions on parameters of voice extractor 26. Feedback also introduces several attractive properties including robustness to errors and disturbances, stability, increased bandwidth, and the like. Feedforward element 50 in feedback voice extractor 26 is represented by **R** which may, in general, represent a matrix or the transfer function of a dynamic model. If the dimensions of **m** and **y** are the same, **R** may be chosen to be the identity matrix. Note that parameter matrices **A**, **B**, **C** and **D** in feedback element 52 do not necessarily correspond with the same parameter matrices in the feedforward system.

The mutual information of a random vector **y** is a measure of dependence among its components and is defined as follows:

$$L(\mathbf{y}) = \sum_{y \in \mathbf{y}} p_y(y) \ln \left| \frac{p_y(y)}{\prod_{j=1}^{j=r} p_{y_j}(y_j)} \right|$$

An approximation of the discrete case is as follows:

$$L(\mathbf{y}) \cong \sum_{k=k_0}^{k_1} p_y(y(k)) \ln \left| \frac{p_y(y(k))}{\prod_{j=1}^{j=r} p_{y_j}(y_j(k))} \right|$$

where $p_y(y)$ is the probability density function of the random vector **y** and $p_{y_j}(y_j)$ is the probability density of the j^{th} component of the output vector **y**. The functional $L(\mathbf{y})$ is always non-negative and is zero if and only if the components of the random vector **y** are statistically independent. This measure defines the degree of dependence among the components of the signal vector. Therefore, it represents an

appropriate function for characterizing a degree of statistical independence. $L(\mathbf{y})$ can be expressed in terms of the entropy:

$$L(\mathbf{y}) = -H(\mathbf{y}) + \sum_i H(y_i)$$

where $H(\cdot)$ is the entropy of $\mathbf{y}$ defined as $H(\mathbf{y}) = -\mathrm{E}[\ln f_y]$ and $\mathrm{E}[\cdot]$ denotes the expected value.

Mixing environment 42 can be modeled as the following nonlinear discrete-time dynamic (forward) processing model:

$$X_p(k+1) = f_p^k(X_p(k), s(k), w_1^*)$$
$$m(k) = g_p^k(X_p(k), s(k), w_2^*)$$

where $s(k)$ is an n-dimensional vector of original sources, $m(k)$ is the m-dimensional vector of measurements and $X_p(k)$ is the N_p-dimensional state vector. The vector (or matrix) w_1^* represents constants or parameters of the dynamic equation and w_2^* represents constants or /parameters of the output equation. The functions $f_p(\cdot)$ and $g_p(\cdot)$ are differentiable. It is also assumed that existence and uniqueness of solutions of the differential equation are satisfied for each set of initial conditions $X_p(t_0)$ and a given waveform vector $s(k)$.

Voice extractor 26 may be represented by a dynamic feedforward network or a dynamic feedback network. The feedforward network is:

$$X(k+1) = f^k(X(k), m(k), w1)$$
$$y(k) = g^k(X(k), m(k), w2)$$

where k is the index, $m(k)$ is the m-dimensional measurement, $y(k)$ is the r-dimensional output vector, and $X(k)$ is the N-dimensional state vector. Note that N and N_p may be different. The vector (or matrix) w_1 represents the parameter of the dynamic equation and the vector (or matrix) w_2 represents the parameter of the output equation. The functions $f(\cdot)$ and $g(\cdot)$ are differentiable. It is also assumed that existence and uniqueness of solutions of the differential equation are satisfied

for each set of initial conditions $X(t_0)$ and a given measurement waveform vector $m(k)$.

The update law for dynamic environments is used to recover the original signals. Environment 42 is modeled as a linear dynamical system. Consequently, voice extractor 26 will also be modeled as a linear dynamical system.

In the case where voice extractor 26 is a feedforward dynamical system, the performance index may be defined as follows:

$$J_0(w_1, w_2) = \sum_{k=k_0}^{k_1-1} L^k(y_k)$$

subject to the discrete-time nonlinear dynamic network

$$X_{k+1} = f^k(X_k, m_k, w_1), \qquad X_{k_0}$$
$$y_k = g^k(X_k, m_k, w_2)$$

This form of a general nonlinear time varying discrete dynamic model includes both the special architectures of multilayered recurrent and feedforward neural networks with any size and any number of layers. It is more compact, mathematically, to discuss this general case. It will be recognized by one of ordinary skill in the art that it may be directly and straightforwardly applied to feedforward and recurrent (feedback) models.

The augmented cost function to be optimized becomes:

$$J_0'(w_1, w_2) = \sum_{k=k_0}^{k_1-1} L^k(y_k) + \lambda_{k+1}^T (f^k(X_k, m_k, w_1) - X_{k+1})$$

The Hamiltonian is then defined as:

$$H^k = L^k(y(k)) + \lambda_{k+1}^T f^k(X, m, w_1)$$

Consequently, the necessary conditions for optimality are:

$$X_{k+1} = \frac{\partial H^k}{\partial \lambda_{k+1}} = f^k(X_k, m_k, w_1)$$

$$\lambda_k = \frac{\partial H^k}{\partial X_k} = (f_{X_k}^k)^T \lambda_{k+1} + \frac{\partial L^k}{\partial X_k}$$

$$\Delta w_2 = -\eta \frac{\partial H^k}{\partial w_2} = -\eta \frac{\partial L^k}{\partial w_2}$$

$$\Delta w_1 = -\eta \frac{\partial H^k}{\partial w_1} = -\eta (f_{w_1}^k)^T \lambda_{k+1}$$

5

The boundary conditions are as follows. The first equation, the state equation, uses an initial condition, while the second equation, the co-state equation, uses a final condition equal to zero. The parameter equations use initial values with small norm which may be chosen randomly or from a given set.

10 In the general discrete linear dynamic case, the update law is then expressed as follows:

$$X_{k+1} = \frac{\partial H^k}{\partial \lambda_{k+1}} = f^k(X, m, w_1) = AX_k + Bm_k$$

$$\lambda_k = \frac{\partial H^k}{\partial X_k} = (f_{X_k}^k)^T \lambda_{k+1} + \frac{\partial L^k}{\partial X_k} = A_k^T \lambda_k + C_k^T \frac{\partial L^k}{\partial y_k}$$

$$\Delta C = -\eta \frac{\partial H^k}{\partial C} = -\eta \frac{\partial L^k}{\partial C} = \eta(-f_a(y)X^T)$$

$$\Delta D = -\eta \frac{\partial H^k}{\partial D} = -\eta \frac{\partial L^k}{\partial D} = \eta([D]^{-T} - f_a(y)m^T)$$

$$\Delta B = -\eta \frac{\partial H^k}{\partial B} = -\eta(f_B^k)^T \lambda_{k+1} = -\eta \lambda_{k+1} m_k^T$$

$$\Delta A = -\eta \frac{\partial H^k}{\partial A} = -\eta(f_A^k)^T \lambda_{k+1} = -\eta \lambda_{k+1} X_k^T$$

The general discrete-time linear dynamics of the network are given as:

$$X(k+1) = A\,X(k) + Bm(k)$$
$$y(k) = C\,X(k) + Dm(k)$$

where $m(k)$ is the m-dimensional vector of measurements, $y(k)$ is the n-dimensional vector of processed outputs, and $X(k)$ is the (mL) dimensional states (representing filtered versions of the measurements in this case). One may view the state vector as composed of the L m-dimensional state vectors $X_1, X_2, \ldots, X_L$. That is,

$$X_k = X(k) = \begin{bmatrix} X_1(k) \\ X_2(k) \\ \cdots \\ X_L(k) \end{bmatrix}$$

In the case where the matrices and A and B are in the controllable canonical form, the A and B block matrices may be represented as:

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1L} \\ I & 0 & \cdots & 0 \\ \cdots & I & \cdots & 0 \\ 0 & 0 & I & 0 \end{bmatrix}, \text{ and } B = \begin{bmatrix} I \\ 0 \\ \cdots \\ 0 \end{bmatrix}$$

where each block sub-matrix A_{ij} may be simplified to a diagonal matrix, and each I is a block identity matrix with appropriate dimensions.

Then:

$$X_1(k+1) = \sum_{j=1}^{L} A_{1j} X_j(k) + m(k)$$

$$X_2(k+1) = X_1(k)$$

$$\cdots$$

$$X_L(k+1) = X_{L-1}(k)$$

$$y(k) = \sum_{j=1}^{L} C_j X_j(k) + Dm(k)$$

This model represents an IIR filtering structure of the measurement vector *m(k)*. In the event that the block matrices A_{ij} are zero, the model is reduced to the special case of an FIR filter.

$$X_1(k+1) = m(k)$$
$$X_2(k+1) = X_1(k)$$
$$\cdots$$
$$X_L(k+1) = X_{L-1}(k)$$

$$y(k) = \sum_{j=1}^{L} C_j X_j(k) + Dm(k)$$

The equations may be rewritten in the well-known FIR form:

$$X_1(k) = m(k-1)$$
$$X_2(k) = X_1(k-1) = m(k-2)$$
$$\cdots$$
$$X_L(k) = X_{L-1}(k-1) = m(k-L)$$

$$y(k) = \sum_{j=1}^{L} C_j X_j(k) + Dm(k)$$

This equation relates the measured signal $m(k)$ and its delayed versions represented by $X_j(k)$, to the output $y(k)$.

The matrices A and B are best represented in the controllable canonical forms or the form I format. Then B is constant and A has only the first block rows as parameters in the IIR network case. Thus, no update equations for the matrix B are used and only the first block rows of the matrix A are updated. Thus, the update law for the matrix A is as follows:

$$\Delta A_{1j} = -\eta \frac{\partial H^k}{\partial A_{1j}} = -\eta (f^k_{A_{1j}})^T \lambda_{k+1} = -\eta \lambda_1(k+1) X_j^T(k)$$

Noting the form of the matrix A, the co-state equations can be expanded as:

$$\lambda_1(k) = \lambda_2(k+1) + C_1^T \frac{\partial L^k}{\partial y_k}(k)$$

$$\lambda_2(k) = \lambda_3(k+1) + C_2^T \frac{\partial L^k}{\partial y_k}(k)$$

$$\vdots$$

$$\lambda_L(k) = C_L^T \frac{\partial L^K}{\partial y_k}(k)$$

$$\lambda_1(k+1) = \sum_{l=1}^{L} C_l^T \frac{\partial L^k}{\partial y_k}(k+l)$$

Therefore, the update law for the block sub-matrices in A are:

$$\Delta A_{1j} = -\eta \frac{\partial H^k}{\partial A_{1j}} = -\eta \lambda_1(k+1) X_j^T(K) = -\eta \sum_{l=1}^{L} C_l^T \frac{\partial l^k}{\partial y_k}(k+l) X_j^T$$

The update laws for the matrices D and C can be expressed as follows:

$$\Delta D = \eta([D]^{-T} - f_a(y)m^T) = \eta(I - f_a(y)(Dm)^T)[D]^{-T}$$

where I is a matrix composed of the r×r identity matrix augmented by additional zero row (if n>r) or additional zero columns (if n<r) and $[D]^{-T}$ represents the transpose of the pseudo-inverse of the D matrix.

For the C matrix, the update equations can be written for each block matrix as follows:

$$\Delta C_j = -\eta \frac{\partial H^k}{\partial C_j} = -\eta \frac{\partial L^k}{\partial C_j} = \eta(-f_a(y)X_j^T)$$

Other forms of these update equations may use the natural gradient to render different representations. In this case, no inverse of the D matrix is used. However, the update law for ΔC becomes more computationally demanding.

If the state space is reduced by eliminating the internal state, the system reduces to a static environment where:

$$m(t) = \overline{D}S(t)$$

In discrete notation, the environment is defined by:

$$m(k) = \overline{D}S(k)$$

Two types of discrete networks have been described for separation of statically mixed signals. These are the feedforward network, where the separated signals $y(k)$ 46 are

$$y(k) = WM(k)$$

and feedback network, where $y(k)$ 46 is defined as:

$$y(k) = m(k) - Dy(k)$$
$$y(k) = (I + D)^{-1} m(k)$$

In case of the feedforward network, the discrete update laws are as follows:

$$W^{t+1} = W^1 + \mu\left\{- f(y(k)) \quad g^T(y(k)) + \alpha I\right\}$$

and in case of the feedback network,

$$D^{t+1} = D^t + \mu\left\{f(y(k))g^T(y(k)) - \alpha I\right\}$$

where (αI) may be replaced by time windowed averages of the diagonals of the $f(y(k))\ g^T(y(k))$ matrix. Multiplicative weights may also be used in the update.

Output separated signals $y(k)$ 46 represent signal sources $s(k)$ 40. As such, at least one component of vector $y(k)$ 46 is extracted speech signal 28 and at least one component of vector $y(k)$ 46 is extracted noise signal 30. Many extracted

speech signals 28 may be simultaneously generated by voice extractor 26. Speech detector 32 may treat each of these as a signal of interest and the remaining as extracted noise signals 30 to generate a plurality of detected speech signals 24.

Referring now to Figure 5, a block diagram illustrating a two transducer voice extractor having a plurality of extracted speech signal outputs according to an embodiment of the present is shown. First extracted speech signal 60 and extracted noise signal 30 provide inputs for voice extract system 62. Voice extract system 62 uses inter-microphone differential information and the statistical properties of independent signal sources to distinguish between audio signals. Algorithms used embody multiple nonlinear mathematical equations that capture the non-linear characteristics and inherent ambiguity in distinguishing between mixed signals in real environments.

Voice extract system 62 generates first output 64 and second output 66. Summer 68 combines sound signal 24 from first microphone (m_1) 22 and second output 66 to produce first extracted speech signal 60. Summer 70 combines sound signal 24 from second microphone (m_2) 22 with first output 64 to generate extracted noise signal 30.

Three other extracted speech signals 28 are also provided. Second extracted speech signal 72 is generated by summer 74 as the difference between sound signal 24 from microphone m_2 22 and extracted noise signal 30. To produce third extracted sound signal 76, extracted noise signal 30 is passed through adaptive least-mean-square (LMS) filter 78. Summer 80 generates third extracted sound signal 76 as the difference between sound signal 24 from microphone m_2 22 and filtered extracted noise signal 82. Similarly, fourth extracted sound signal 84 is based on extracted noise signal 30 filtered by adaptive LMS filter 86. Summer 88 generates fourth extracted sound signal 84 as the difference between sound signal 24 from microphone m_1 22 and filtered extracted noise signal 90 from adaptive LMS filter 86.

Referring now to Figure 6, a block diagram of a two transducer voice extractor generating one extracted speech signal and one extracted noise signal according to an embodiment of the present invention is shown. First filter (W1) 100 receives sound signal 24 from first microphone 22 and generates first filtered output 102. Similarly, second filter (W2) 104 receives sound signal 24 from second microphone 22 and generates second filtered output 106. Summer 108 subtracts second filtered output from sound signal 24 of first microphone 22 to produce first compensated signal 110. Summer 112 subtracts first filtered output 102 from sound signal 24 of second microphone 22 to produce second compensated signal 114. Static unmixer 116 accepts first compensated signal 110 and second compensated signal 114 and generates extracted speech signal 28 and extracted noise signal 30.

This implementation of voice extraction can be thought of as a means of undoing a mixing, which is not only instantaneous as in

$$mix_i(t) = \sum_{j=1}^{N} a_{ij} signal_j(t)$$

where a_{ij} is an entry of the static mixing matrix A, but also involves delayed versions of the signals which can be expressed mathematically as follows:

$$mix_i(t) = \sum_{j=1}^{N} \int_0^{\infty} a_{ij}(t') \; signal_j(t-t')dt'$$

In discrete interpretation of the above, the mixing matrix A, composed of entries a_{ij}, is no longer a single matrix, but a series of matrices A(t=τ) as follows:

$$\mathbf{mix}(t) = \sum_{\tau=0}^{N} A(\tau) \;\; \mathbf{signal}(t-\tau)$$

where **mix** and **signal** are vectors.

There is an element of instantaneous mixture in this expression, where $\tau=0$, which is undone by static unmixer 116. The delayed elements of the mixings are undone by multitap filters W1 100 and W2 104.

Filter coefficients for W1 100, W2 104, and static unmixer 116 can be obtained adaptively, using a variety of criteria. One such criterion is the statistical independence of independent signal sources principle. However, instead of enforcing the constraint at a single time point (i.e., t=0), the adaptation enforces this criterion for all delayed versions (i.e., $t=\tau$), as well. Voice extraction is thus performed by a feedback architecture that follows the equation:

$$y(t)=StaticUnmixer\{\,[\,mix(t)-\sum_{i=1}^{N} W_i\ y(t-i)\,]\,\}$$

where y(t) is the output vector containing extracted speech signal 28 and extracted noise signal 30, mix(t) is the input vector of sound signals 24, and W_i are delayed tap matrices for filters 100, 104, both having zero-diagonals. The filters W_i 100, 104 subtract off delayed versions of the interfering signals.

Static unmixer 116 can be an operator, which involves a matrix multiplication operation reduced to a filter, such as the following:

$$y(t)=(I+D)^{-1}\ [\ mix(t)-\sum_{i=1}^{N} W_i\ y(t-i)\ ]$$

where I is the identity matrix, D is another matrix with zero diagonals.

Assuming a two-input, two-output system, adaptation of the off-diagonal entries of the 2×2 matrices D and W_i can be defined by the following equations:

$$\Delta D=\eta\begin{bmatrix} 0 & f(y_1(t))\ \ g(y_2(t)) \\ f(y_2(t))\ \ g(y_1(t)) & 0 \end{bmatrix}$$

$$\Delta W_i = \eta \begin{bmatrix} 0 & f(y_1(t)) & g(y_2(t-i)) \\ f(y_2(t)) & g(y_1(t-i)) & 0 \end{bmatrix}$$

where η is the rate of adaptation, $y_i(t)$ is the scalar output y_i at time t, and $f(x)$ and $g(x)$ are functions with certain mathematical properties. As will be recognized by one of ordinary skill in the art, these functions and various filter coefficients depend on a variety of variables, including the type and relative placement of transducers 22, type and level of noise expected, sampling rate, application, and the like.

Referring now to Figure 7, a block diagram illustrating a voice detector according to an embodiment of the present invention is shown. Voice detector 32 includes speech feature extractor 130 receiving one or more extracted speech signals 28 and generating one or more speech signal properties 132. Noise feature extractor 134 receives one or more extracted noise signals 30 and generates one or more noise signal properties 136. As will be described in greater detail below, properties 132, 136 can convey any information about extracted speech signals 28 and extracted noise signals 30, respectively. For example, properties 132, 136 may include one or more of signal powers, statistical properties, spectral properties, envelope properties, proximity between transducers 22, and the like. For example, extracted signals 28, 30 may be smoothed to produce signal envelopes and at least one property extracted from each envelope, such as local peaks or valleys, averages, threshold crossings, statistical properties, model fitting values, and the like. One or more properties used for speech signal property 132 may be the same as or correspond with properties used for noise signal property 136.

Comparor 138 generates at least one detection parameter 140 based on speech signal properties 132 and noise signal properties 136. Comparor 138 may operate in a variety of manners. For example, comparor 138 may generate detection parameter 140 as a mathematical combination of speech signal property 132 and noise signal property 136 such as, for example, a difference or a ratio. The

result of this operation may be output directly as detection parameter 140, may be scaled to produce detection parameter 140, or detection parameter 140 may be a binary value resulting from comparing the operation results to one or more threshold values.

Attenuator 142 attenuates extracted speech signals 28 based on detection parameter 140 to produce detected speech signal 34. Detected speech signal 34 may also include detection parameter 140 as an indication of whether or not speech is present in extracted speech signal 28.

Referring now to Figure 8, a block diagram illustrating a voice detector using multiple frequency bands according to an embodiment of the present invention is shown. Speech detector 32 includes time windower 150 accepting one or more extracted speech signals 28 and producing windowed speech signals 152. Similarly, time windower 154 accepts one or more extracted noise signals 30 and produces windowed noise signals 156. Windowing operations performed by windowers 150, 154 may be overlapping or non-overlapping and may implement a variety of windowing filters such as, for example, Hanning filters, Hamming filters, and the like.

Frequency converter 158 generates speech frequency bands, shown generally by 160, from windowed speech signal 152. Similarly, frequency converter 162 generates noise frequency bands, shown generally by 164, for each windowed noise signal 156. Frequency converters 158, 162 may implement any algorithm which generates spectral information from windowed signals 152, 156, respectively. For example, frequency converter 158, 162 may implement a fast Fourier transform (FFT) algorithm.

For each speech frequency band 160, criteria applier 166 accepts one speech frequency band 160 and a corresponding noise frequency band 164 and generates frequency band output 168 based on at least one detection parameter. Each detection parameter is based on at least one property of speech frequency band 160 and on corresponding noise frequency band 164. Any property of speech

frequency band 160 or noise frequency band 164 may be used. Such properties include in-band power, magnitude properties, phase properties, statistical properties, and the like. For example, frequency band output 168 may be based on the ratio of in-band speech signal power to in-band noise signal power. Frequency band output 168 may include speech frequency band 160 scaled by the ratio of speech in-band power to noise in-band power. Alternatively, frequency band output 168 may attenuate speech frequency band 160 if the in-band signal-to-noise ratio is below a threshold.

Combiner 170 combines frequency band output 168 for each speech frequency band 160 to generate detected speech signal 34. In one embodiment, combiner 170 performs inter-band filtering followed by an inverse-FFT to generate detected speech signal 34. Alternatively or in combination, combiner 170 examines each frequency band output 168 and generates detected speech signal 34 indicating the likelihood that speech is present.

Referring now to Figures 9 and 10, histogram plots of a typical voice signal and a typical noise signal, respectively, are shown. Voice signals tend to have Laplacian probability distribution, such as shown in voice signal histogram plot 180. Noise signals, on the other hand, tend to have a Gaussian or Super-Gaussian probability distribution, such as seen in noise signal histogram plot 182. Thus, voice signals can be said to be of lower variance. The variance of extracted speech signal 28 or speech frequency bands 160 may be used to determine the presence of voice. Various other statistical measures, such as kirtosis, standard deviation, and the like, may be extracted as properties of speech and noise signals or frequency bands.

Referring now to Figures 11 and 12, frequency plots of a typical voice signal and a typical noise signal, respectively, are shown. The spectrum for speech, such as shown by voice power spectral density 190, is different then for noise, shown by noise power spectral density plot 192. Voice signals tend to have a narrower band width with pronounced peaks at formants. In contrast, most noise generally has a broader bandwidth. Various spectral techniques are possible. For

example, one or more estimated bandwidth may be used. Statistical characteristics of the magnitude spectrum may also be extracted. Further, frequency spectrums 190, 192 may be used to derive parameters of a model. These parameters would then serve as signal properties.

Referring now to Figure 13, a schematic diagram illustrating relative transducer placement for a proximity-based speech detection according to an embodiment of the present invention is shown. Sources of voice signals, such as speaker 200, tend to be closer to transducers 22 then noise sources 202. This is true, for example, if user 200 is holding a palm top device at arms length. A microphone 22 on the palm top device is much closer to voice source 200 while one or more interfering noise sources 202 are usually much further away. Other effects of proximity may be evident in the presence of echos. Echos of a signal that is close to transducer 22 will be weaker then echos of sound sources far away. Still other effects of proximity may emerge when more then one transducer 22 are used. For signal sources that are close to multiple transducers 22, the difference in amplitude between transducers 22 will be more pronounced then signals that are further away. The arrangement of transducers 22 may be organized to amplify this effect. For example, two transducers 22 may be aligned with speaker 200 along axis 204. For any noise source 202 off of axis 204, the ratio of path lengths a,b from noise source 202 to transducers 22 will be less then the ratio of path lengths c,d from speaker 200 to transducers 22. This effect is exaggerated by the fact that sound decreases as the square of the distance. Thus, sound signal 24 from microphone 22 closer to speaker 200 is "speechier" and sound signal 24 from microphone 22 farther from speaker 200 is "noisier" by way of the arrangement of microphones 22.

Referring now to Figures 14 and 15, plots of a noisy speech signal and a speech detected signal according to an embodiment of the present invention, respectively, are shown. Noisy signal 210 contains periods of noise information between speech utterances. Speech detected signal 212 has such noisy periods attenuated. Because silence may be coded or compressed at a lower rate then speech, the result may be used to reduce the number of bits needed to be stored or

sent over a channel.

Referring now to Figure 16, compressing or coding according to an embodiment of the present invention is shown. A coder/compressor system, shown generally by 220, includes speech detector 32 generating one or more detected speech signals 34. Detected speech signal 34 includes speech likelihood signal 222 expressing the likelihood that speech is present. Speech likelihood signal 222 may be a binary signal or may express some probability that speech has been detected by speech detector 32.

Coder/compressor 224 accepts speech likelihood signal 222 and generates coded or compressed signal 226 based on speech likelihood signal 222. Coder/compressor 224 also receives speech signal source 228 which may be an output of speech detector 32, extracted speech signal 28, or sound signal 24 from transducer 22. Coder/compressor 224 variably encodes and/or compresses speech signal source 228 based on speech likelihood signal 222. Thus, coded/compressed signal 226 requires substantially fewer bits. This may result in a wide variety of benefits including less bandwidth required, less storage required, greater data accuracy, greater information throughput, and the like.

While embodiments of the invention have been illustrated and described, it is not intended that these embodiments illustrate and describe all possible forms of the invention. The words of the specification are words of description rather then limitation, and it is understood that various changes may be made without departing from the spirit and scope of the invention.

Many embodiments have been shown in block diagram form for ease of illustration. However, one of ordinary skill in the art will recognize that the present invention may be implemented in any combination of hardware and software and in a wide variety of devices such as computers, digital signal processors, custom integrated circuits, programmable logic devices, analog components, and the like. Further, blocks may be logically combined or further subdivided to suit a particular implementation.

WHAT IS CLAIMED IS:

1. A speech detection system comprising:

at least one transducer converting sound into an electrical signal;

a voice extractor in communication with the at least one transducer, the voice extractor producing at least one extracted speech signal and at least one extracted noise signal based on at least one electrical sound signal; and

a speech detector in communication with the voice extractor, the speech detector generating a detected speech signal based on the at least one extracted speech signal and on the at least one extracted noise signal.

2. A speech detection system as in claim 1 wherein the speech detector recognizes periods of speech based on at least one property of the at least one extracted speech signal and on at least one corresponding property of the at least one extracted noise signal.

3. A speech detection system as in claim 1 wherein the speech detector recognizes periods of speech based on statistical properties of the at least one extracted speech signal and on statistical properties of the at least one extracted noise signal.

4. A speech detection system as in claim 1 wherein the speech detector recognizes periods of speech based on spectral properties of the at least one extracted speech signal and on spectral properties of the at least one extracted noise signal.

5. A speech detection system as in claim 1 wherein the at least one transducer is a plurality of transducers, the speech detector recognizing periods of speech based on estimated relative proximity of a speaker to at least two of the plurality of transducers.

6. A speech detection system as in claim 1 wherein the speech detector recognizes periods of speech based on an envelope of the at least one extracted speech signal.

7. A speech detection system as in claim 1 wherein the at least one extracted speech signal is divided in time into a plurality of windows, the speech detector generating the detected speech signal based on determining whether or not speech is present in each window.

8. A speech detection system as in claim 7 wherein the at least one extracted speech signal is divided into a plurality of frequency bands, the speech detector determining whether or not speech is present in each frequency band for each window.

9. A speech detection system as in claim 8 wherein the detected speech signal is based on combining the determination for each frequency band for each window.

10. A speech detection system as in claim 1 further comprising a variable rate coder in communication with the speech detector, the variable rate coder changing a coding rate for coding the detected speech signal based on a determined presence of speech in the detected speech signal.

11. A speech detection system as in claim 1 further comprising a variable rate compressor in communication with the speech detector, the variable rate compressor changing a compression rate for compressing the detected speech signal based on a determined presence of speech in the detected speech signal.

12. A method of detecting speech in the presence of noise comprising:

receiving at least one signal containing speech mixed with noise;

extracting at least one extracted speech signal from the at least one received signal;

extracting at least one extracted noise signal from the at least one received signal; and

generating a detected speech signal based on the at least one extracted speech signal and the at least one extracted noise signal.

13. A method of detecting speech as in claim 12 wherein the detected speech signal comprises periods wherein the at least one extracted speech signal is attenuated.

14. A method of detecting speech as in claim 12 wherein the detected speech signal comprises a likelihood of speech presence.

15. A method of detecting speech as in claim 12 wherein generating the detected speech signal comprises comparing at least one statistical property from the at least one extracted speech signal with at least one corresponding statistical property from the at least one extracted noise signal.

16. A method of detecting speech as in claim 12 wherein generating the detected speech signal comprises comparing at least one spectral property from the at least one extracted speech signal with at least one corresponding spectral property from the at least one extracted noise signal.

17. A method of detecting speech as in claim 12 wherein receiving at least one signal comprises receiving one signal from each of a plurality of acoustic transducers.

18. A method of detecting speech as in claim 17 wherein generating the detected speech signal is based on relative proximities to a speaker of at least two of the acoustic transducers.

19. A method of detecting speech as in claim 12 wherein generating the detected speech signal comprises comparing at least one envelope property from the at least one extracted speech signal with at least one corresponding envelope property from the at least one extracted noise signal.

20. A method of detecting speech as in claim 12 further comprising dividing the at least one extracted speech signal in time into a plurality of windows, the speech detector generating a detected speech signal based on determining whether or not speech is present in each window.

21. A method of detecting speech as in claim 20 further comprising dividing the at least one extracted speech signal into a plurality of frequency bands, wherein generating a detected speech signal comprises determining whether or not speech is present in each frequency band.

22. A method of detecting speech as in claim 21 wherein generating the detected speech signal further comprises combining the determination for each frequency band for each window.

23. A method of detecting speech as in claim 12 further comprising determining a coding rate based on a determined presence of speech in the detected speech signal.

24. A method of detecting speech as in claim 12 further comprising determining a compression rate based on a determined presence of speech in the detected speech signal.

25. A method of detecting speech as in claim 12 wherein generating the detected speech signal comprises comparing at least one property of the extracted speech signal with at least one corresponding property of the at least one extracted noise signal.

26. A method of detecting speech comprising:

receiving at least one noise signal;

receiving at least one speech signal having a greater content of the speech than the at least one noise signal;

extracting at least one noise parameter from the at least one noise signal;

extracting at least one speech parameter from the at least one speech signal;

comparing the at least one speech parameter and the at least one noise parameter; and

detecting the presence of speech based on the comparison.

27. A method of detecting speech as in claim 26 wherein extracting at least one noise parameter comprises time windowing the received at least one noise signal and wherein extracting at least one speech parameter comprises time windowing the received at least one speech signal.

28. A method of detecting speech as in claim 27 wherein extracting at least one noise parameter comprises dividing the windowed at least one noise signal into a first plurality of frequency bands and wherein extracting at least one speech parameter comprises dividing the at least one windowed speech signal into second plurality of frequency bands.

29. A method of detecting speech as in claim 28 wherein comparing comprises comparing each noise signal frequency band with a corresponding speech signal frequency band.

30. A method of detecting speech as in claim 29 wherein detecting the presence of speech comprises detecting the presence of speech for each frequency band.

31. A method of detecting speech comprising:

receiving a noise signal;

receiving a speech signal having greater speech content than the noise signal;

dividing the speech signal into a plurality of speech frequency bands;

dividing the noise signal into a plurality of noise frequency bands, each noise frequency band corresponding to one of the speech frequency bands;

for each speech frequency band, calculating at least one detection parameter based on at least one property of the speech frequency band and on at least one property of the corresponding noise frequency band;

for each speech frequency band, generating a frequency band output based on the at least one detection parameter for the speech frequency band.

32. A method of detecting speech as in claim 31 wherein the at least one property of the speech frequency band comprises speech power in the speech frequency band and wherein the at least one property of the noise frequency band comprises noise power in the noise frequency band.

33. A method of detecting speech as in claim 32 wherein calculating at least one detection parameter for each speech frequency band comprises calculating a ratio of speech power in the speech frequency band to noise power in the corresponding noise frequency band.

34. A method of detecting speech as in claim 31 wherein generating a frequency band output comprises attenuating the speech frequency band based on the at least one detection parameter for the speech frequency band.

35. A method of detecting speech as in claim 31 further comprising combining the frequency band output for each speech frequency band.

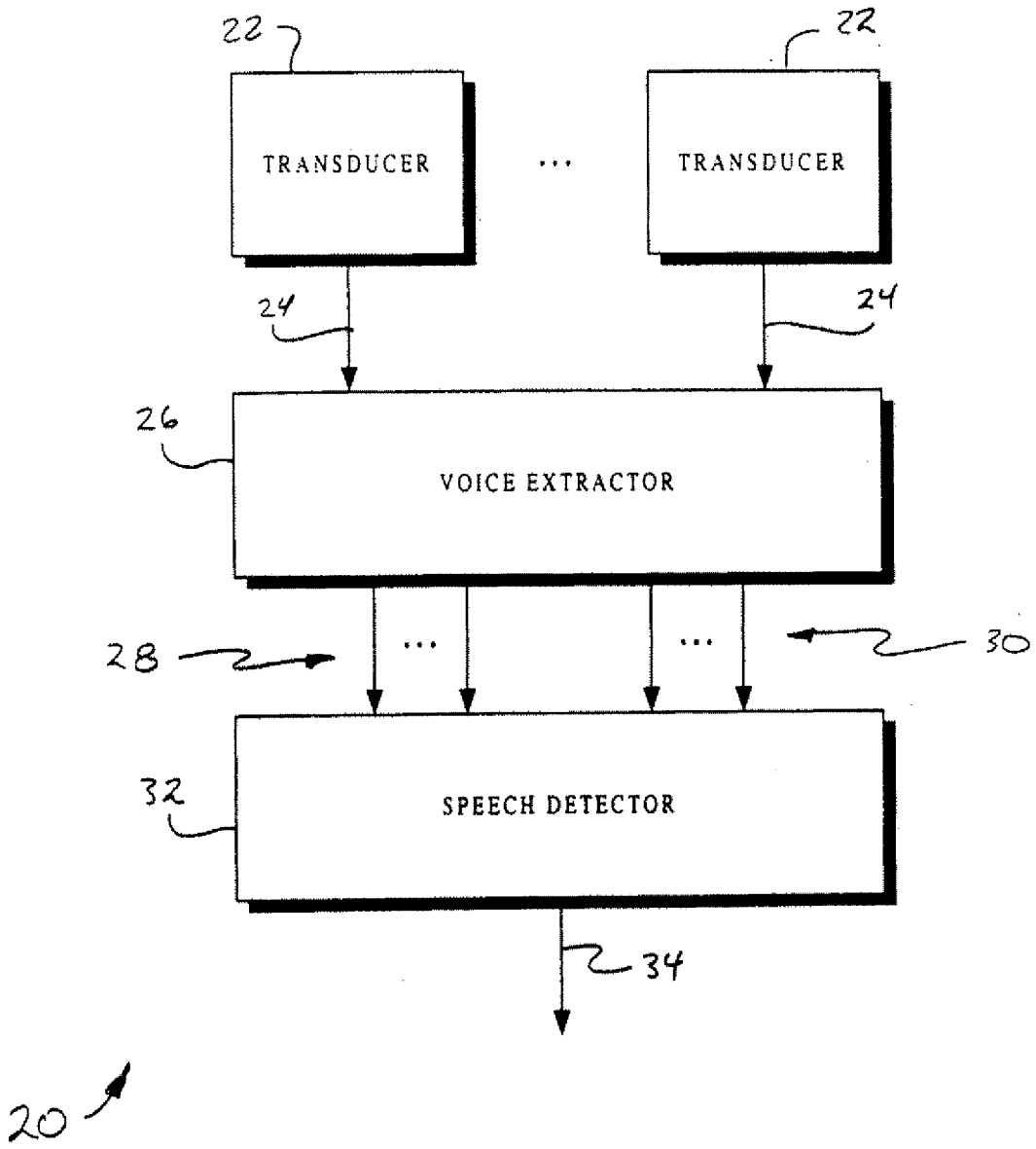

22
22
TRANSDUCER
TRANSDUCER
24
24
26
VOICE EXTRACTOR
28
30
32
SPEECH DETECTOR
34
20

FIG. 1

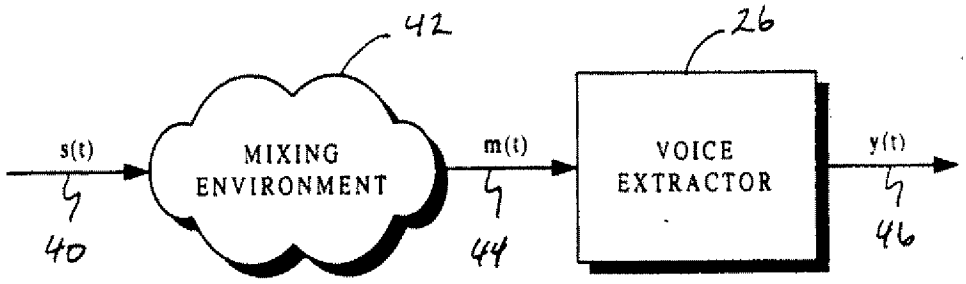


FIG. 2

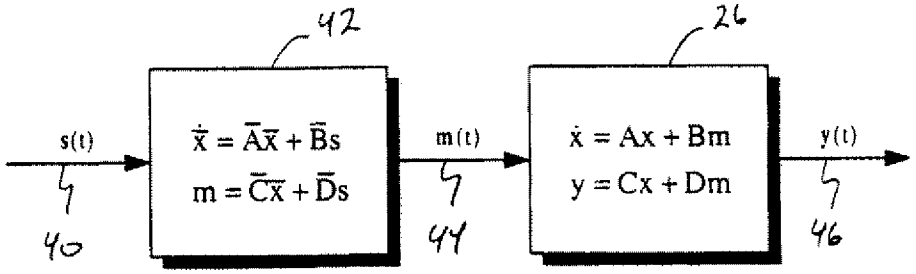

42
26
s(t)
$\dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}s$
$m = \bar{C}\bar{x} + \bar{D}s$
m(t)
$\dot{x} = Ax + Bm$
$y = Cx + Dm$
y(t)
40
44
46

FIG. 3

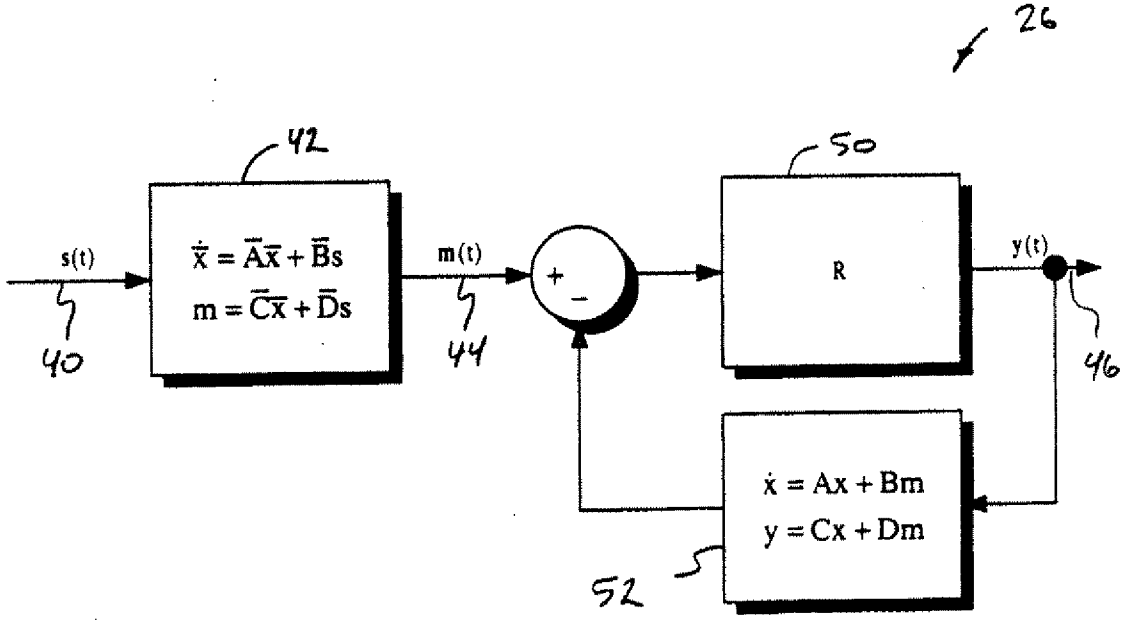

26
42
50
s(t)
$\dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}s$
$m = \bar{C}\bar{x} + \bar{D}s$
m(t)
+
−
R
y(t)
$\dot{x} = Ax + Bm$
$y = Cx + Dm$
40
44
46
52

FIG. 4

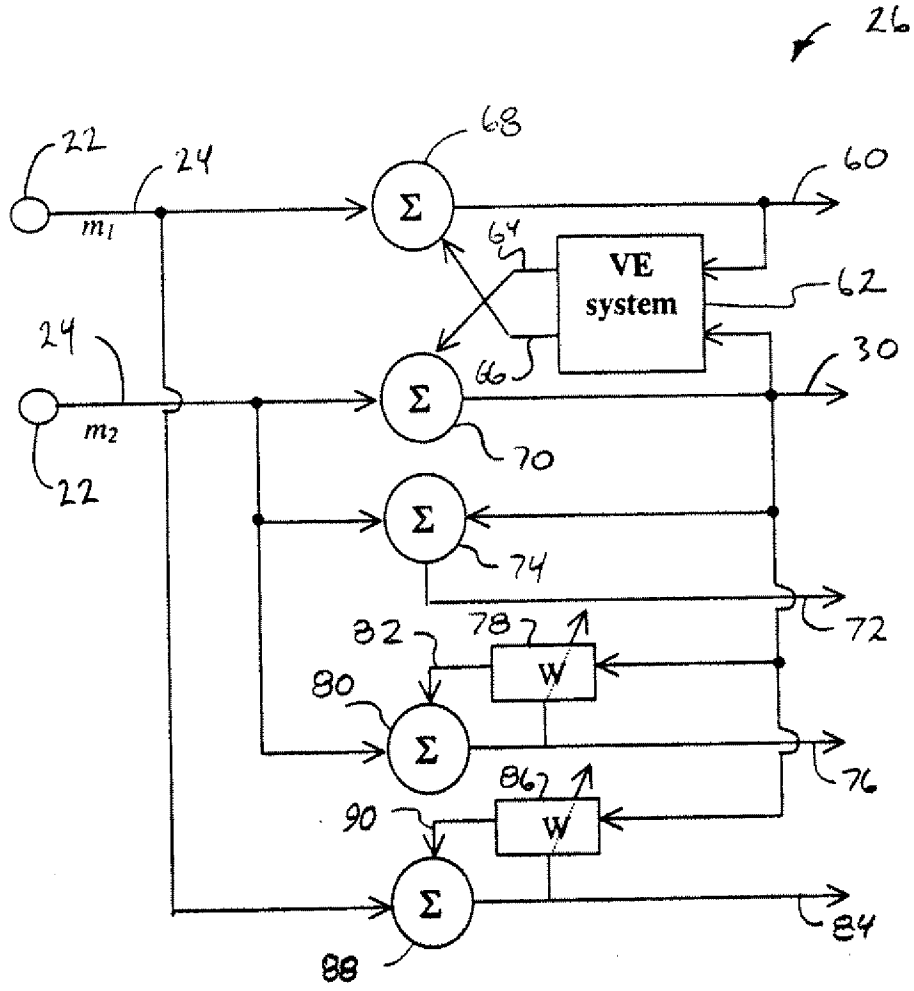

26
22
24
m_1
68
60
64
VE system
62
24
66
30
m_2
22
70
74
72
82
78
W
80
76
86
W
90
84
88

FIG. 5

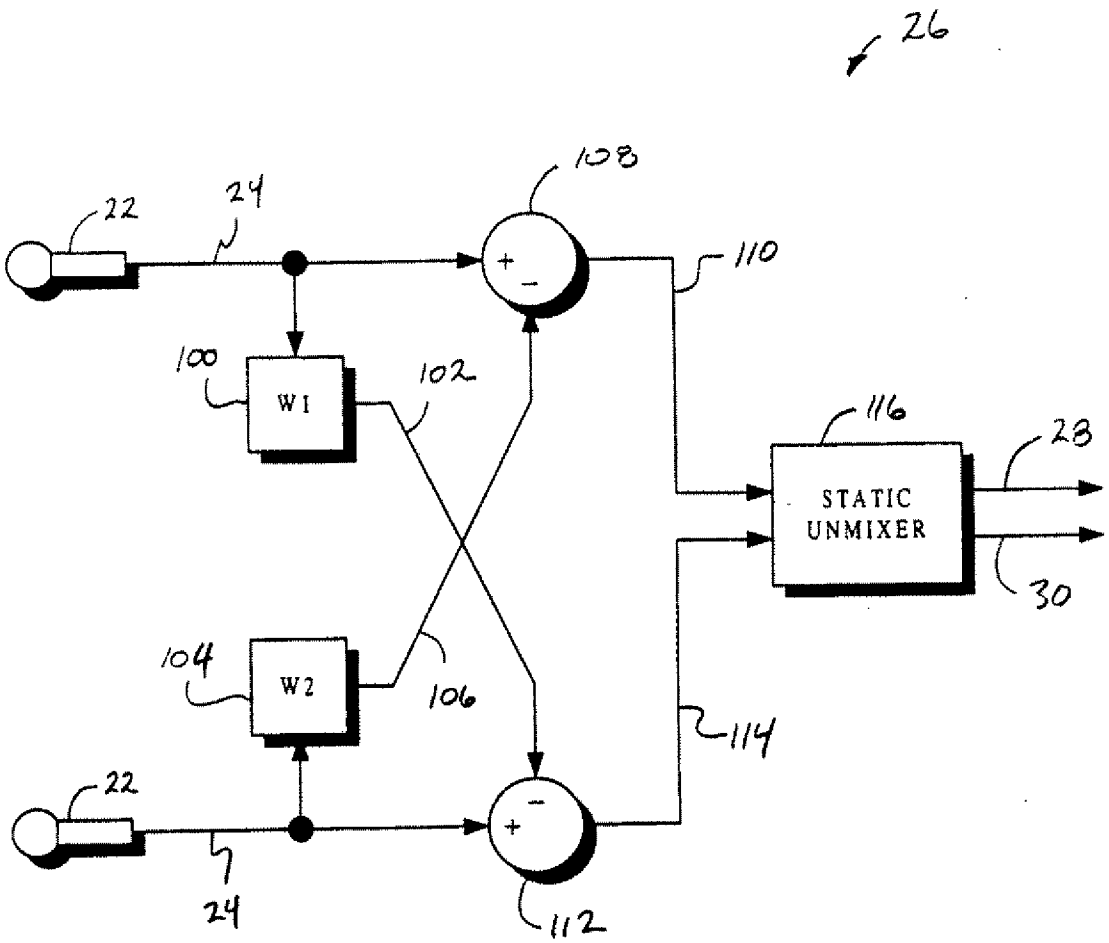

26
22
24
108
110
100
W1
102
116
28
STATIC
UNMIXER
30
104
W2
106
114
22
24
112

FIG. 6

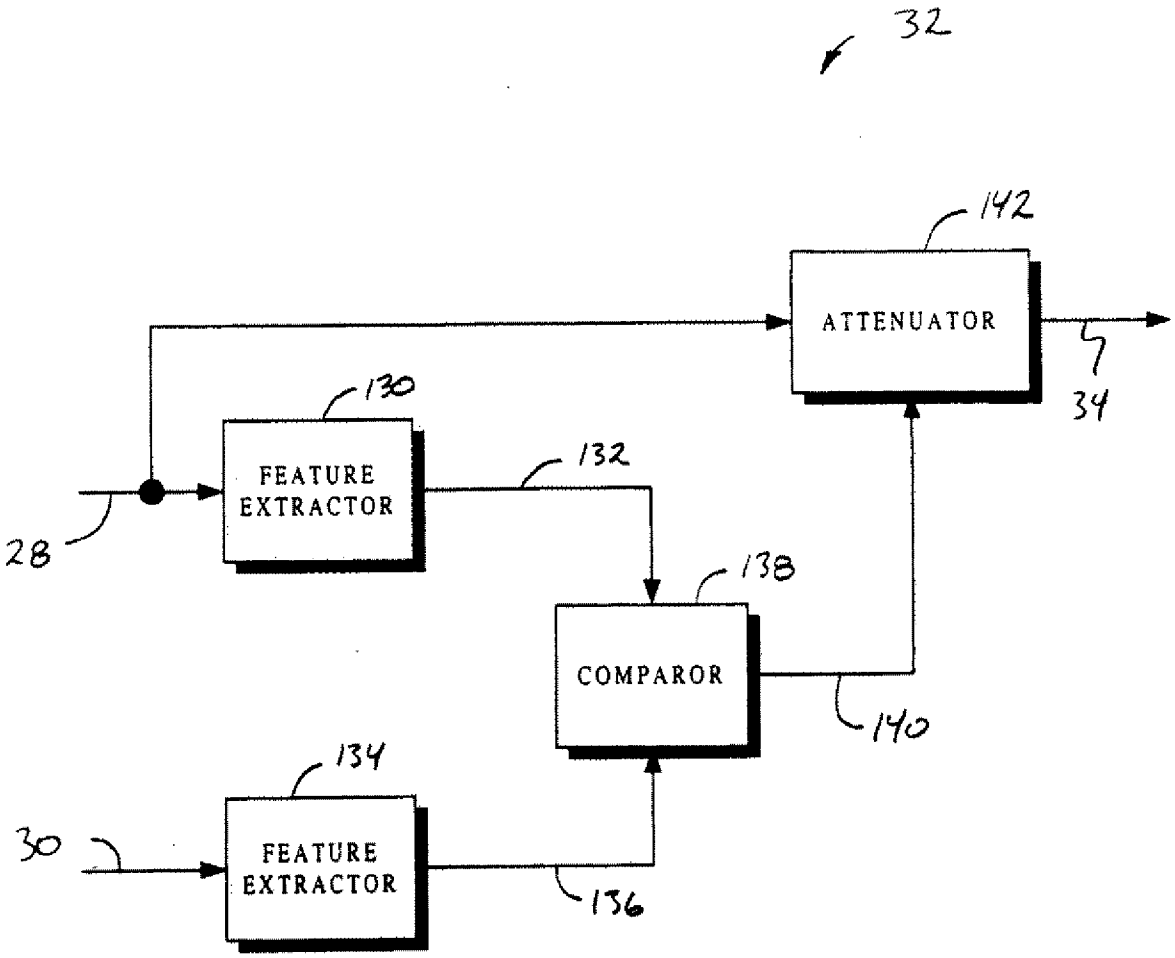

32
142
ATTENUATOR
34
130
FEATURE EXTRACTOR
132
28
138
COMPARATOR
140
134
30
FEATURE EXTRACTOR
136

FIG. 7

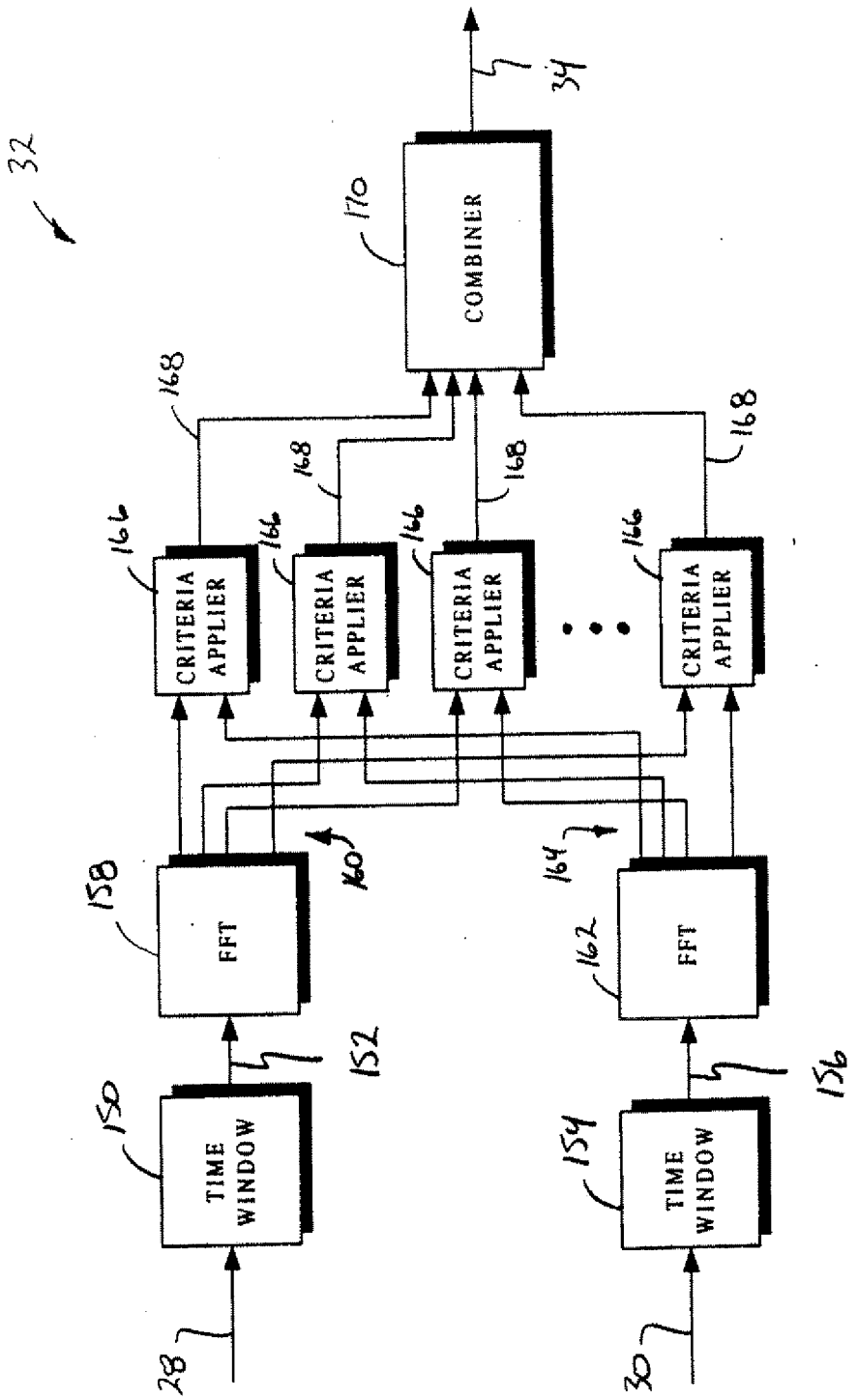

32
34
170
COMBINER
168
168
168
168
166
166
166
166
CRITERIA APPLIER
CRITERIA APPLIER
CRITERIA APPLIER
CRITERIA APPLIER
158
160
164
FFT
FFT
162
152
156
150
154
TIME WINDOW
TIME WINDOW
28
30

FIG. 8

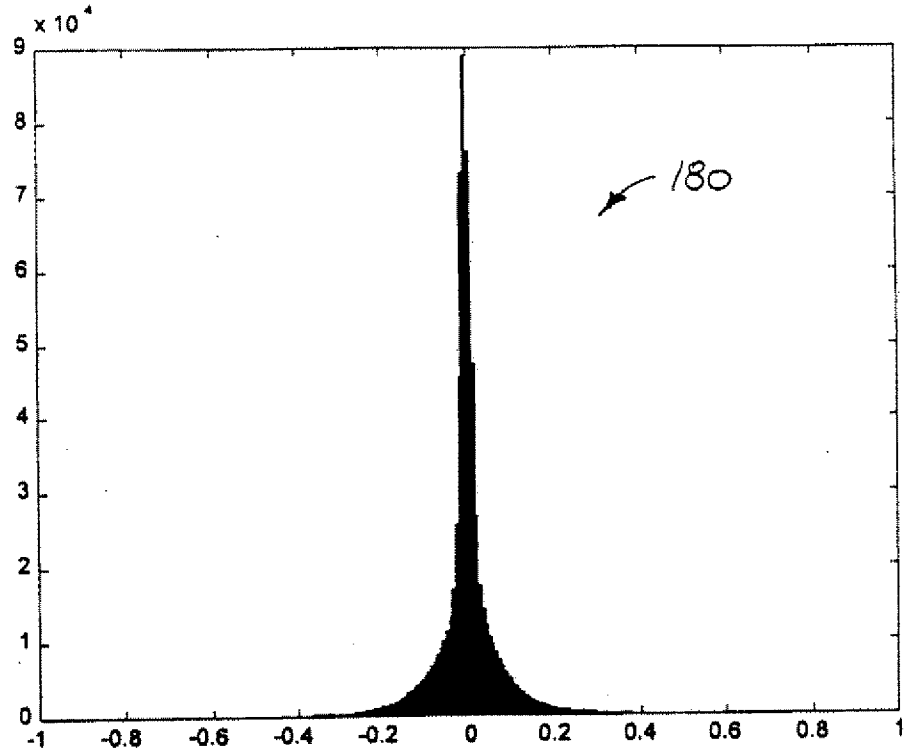


FIG. 9

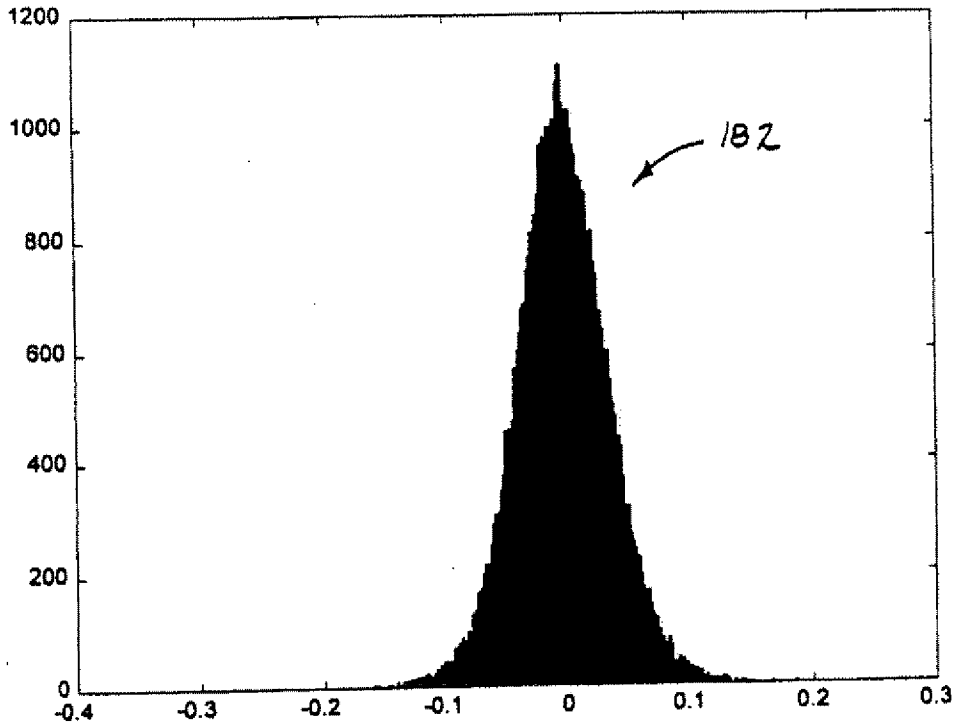


FIG. 10

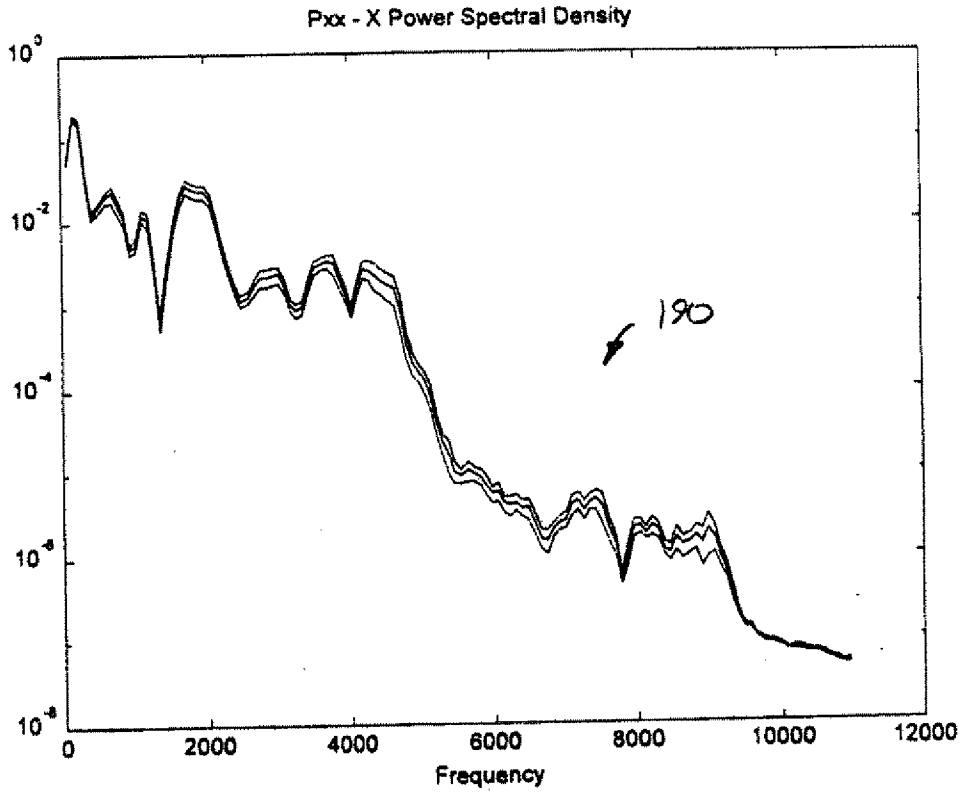

Pxx - X Power Spectral Density
10^0
10^-2
10^-4
10^-6
10^-8
0
2000
4000
6000
8000
10000
12000
Frequency
190

FIG. 11

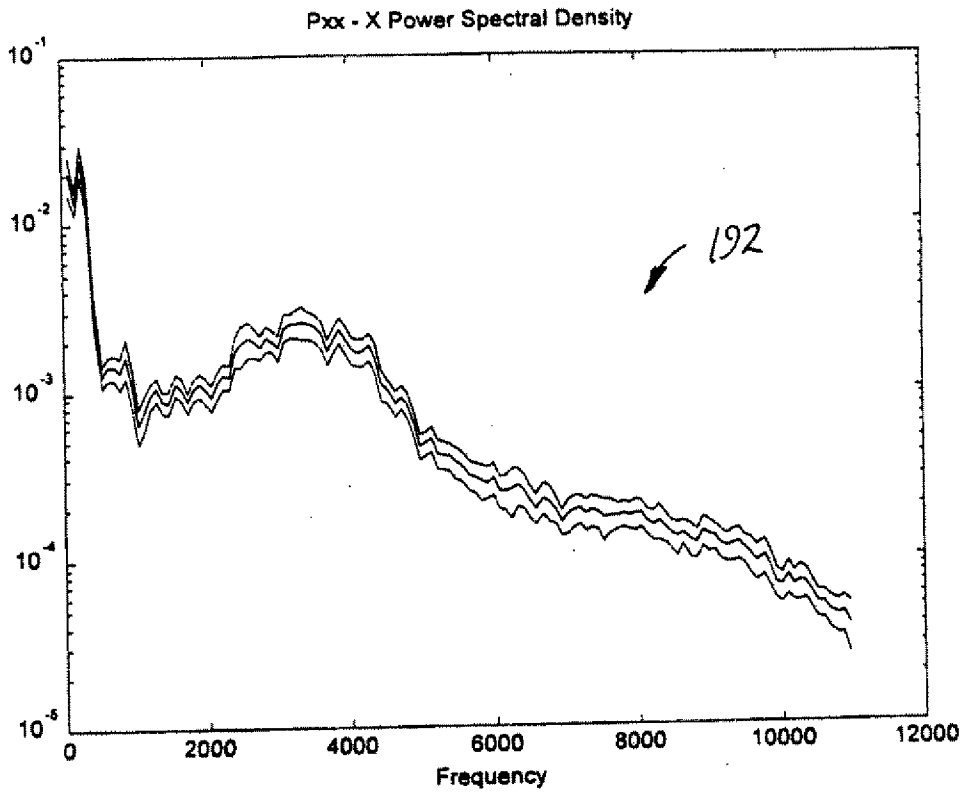
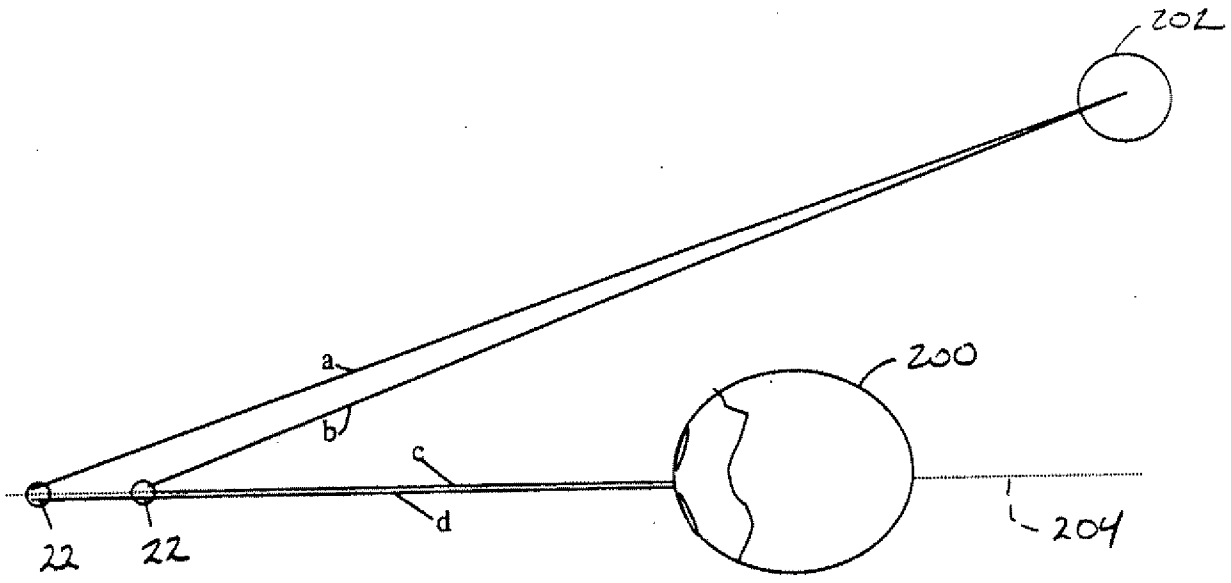

Pxx - X Power Spectral Density
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
0
2000
4000
6000
8000
10000
12000
Frequency
192

FIG. 12



$$\frac{a}{b} < \frac{d}{c}$$

FIG. 13

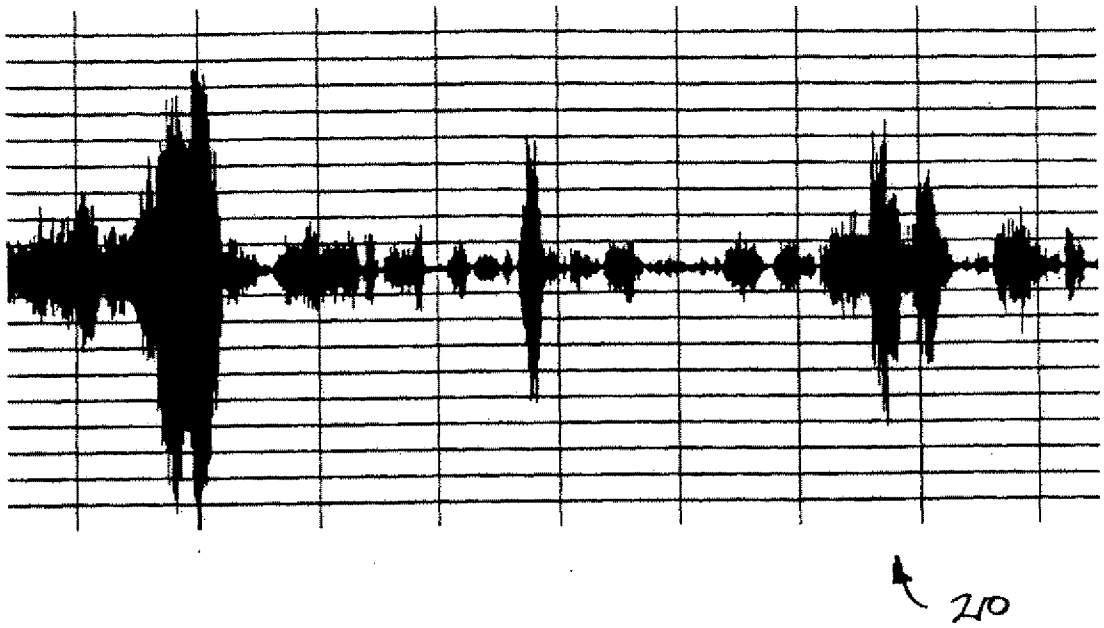


FIG. 14

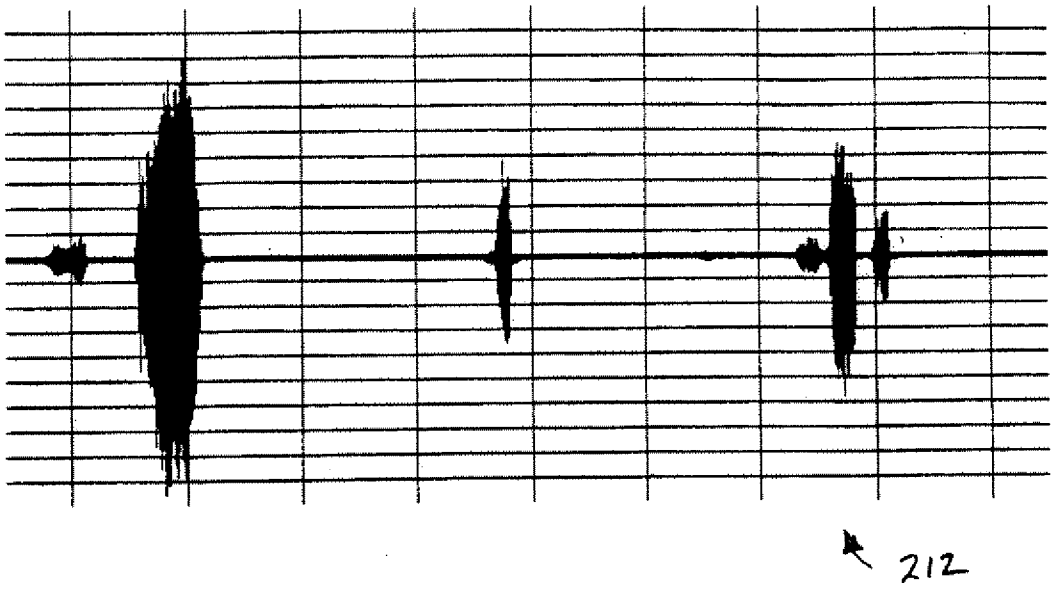


FIG. 15

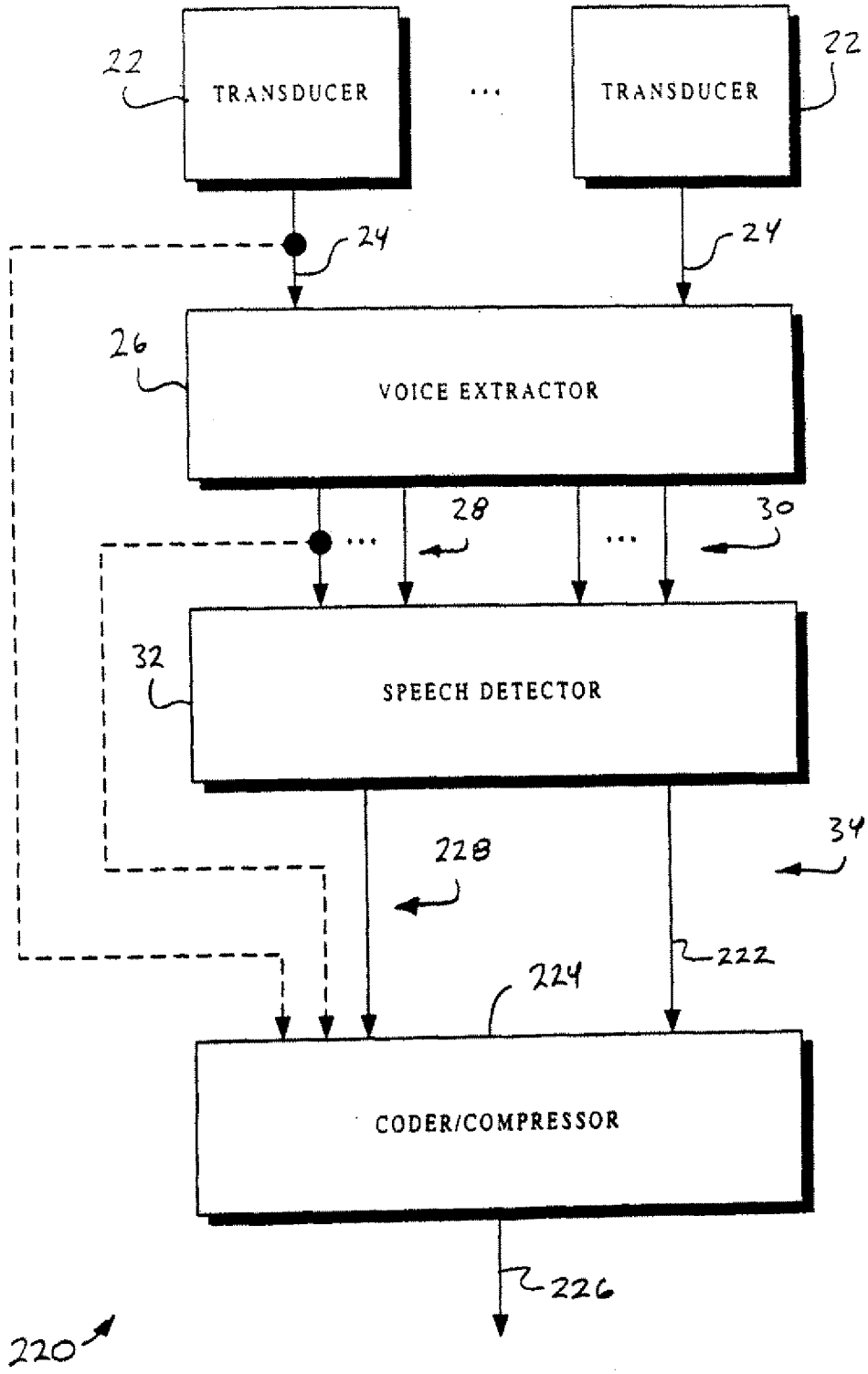

22
TRANSDUCER
...
TRANSDUCER
22
24
24
26
VOICE EXTRACTOR
28
30
32
SPEECH DETECTOR
228
34
222
224
CODER/COMPRESSOR
226
220

FIG. 16

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 April 2002 (11.04.2002)

PCT

(10) International Publication Number
WO 02/29780 A3

(51) International Patent Classification[7]: G10L 11/02, 21/02

(21) International Application Number: PCT/US01/31121

(22) International Filing Date: 3 October 2001 (03.10.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/238,560 4 October 2000 (04.10.2000) US

(71) Applicant *(for all designated States except US)*: **CLARITY, LLC** [US/US]; 3290 West Big Beaver Road, Suite 200, Troy, MI 48084 (US).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **ERTEN, Gamze** [US/US]; 1848 Elk Lane, Okemos, MI 48864 (US).

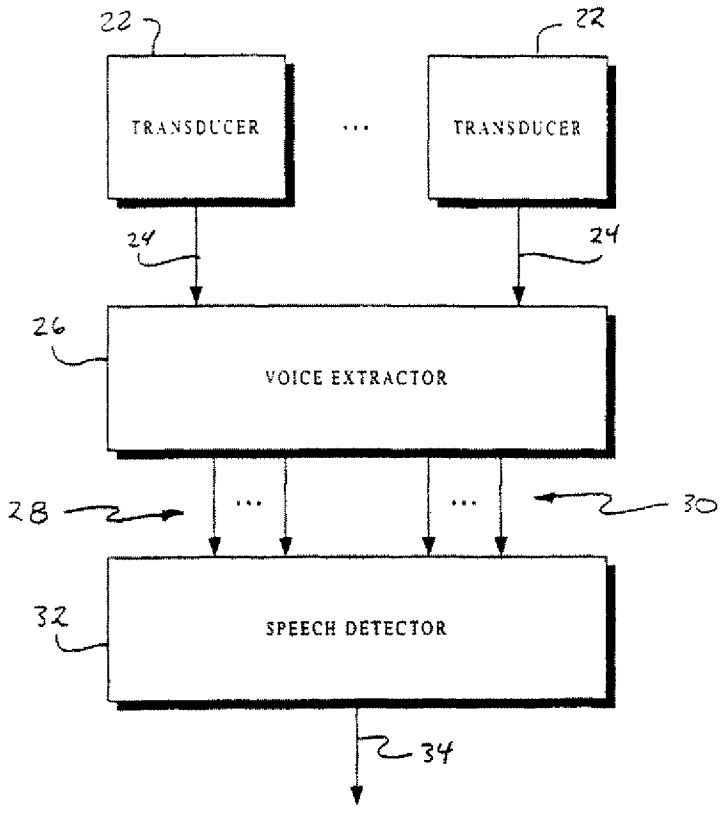
(74) Agents: **CHUEY, Mark, D.** et al.; Brooks & Kushman, 1000 Town Center, Twenty-Second Floor, Southfield, MI 48075 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian

[Continued on next page]

(54) Title: SPEECH DETECTION WITH SOURCE SEPARATION



(57) Abstract: Speech in the presence of noise is detected by first extracting at least one extracted speech signal (28) from at least one received signal (24) and extracting at least one extracted noise signal (30) from the at least one received signal (24). A detected speech signal (34) is generated based on both at least one extracted speech signal (28) and on at least one extracted noise signal (30).

WO 02/29780 A3

patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

(88) Date of publication of the international search report:
20 June 2002

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Int tional Application No
PCT/US 01/31121

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L11/02 G10L21/02

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
INSPEC, EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	PAOLETTI D R ET AL: "Enhanced silence detection in variable rate coding systems using voice extraction" PROCEEDINGS OF THE 43RD IEEE MIDWEST SYMPOSIUM ON CIRCUITS AND SYSTEMS (CAT.NO.CH37144), LANSING, MI, USA, , 8 - 11 August 2000, pages 592-594 vol.2, XP002194768 2000, Piscataway, NJ, USA, IEEE, USA ISBN: 0-7803-6475-9	1,7, 10-12, 20,23-27
Y	the whole document	3-6,8,9, 13-19, 21,22, 28-35

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
2 April 2002	15/04/2002

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Quélavoine, R

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5 630 015 A (KANE ET AL) 13 May 1997 (1997-05-13) abstract; figures 3,7,8,17 column 5, line 64-67 column 9, line 8-12	3,4,6,8, 9,13-16, 19,21, 22,28-35
Y	NAKATANI T ET AL: "Harmonic sound stream segregation using localization and its application to speech stream segregation" SPEECH COMMUNICATION, ELSEVIER SCIENCE PUBLISHERS, AMSTERDAM, NL, vol. 27, no. 3-4, April 1999 (1999-04), pages 209-222, XP004163251 ISSN: 0167-6393 abstract	5,17,18
A	ERTEN G ET AL: "VOICE EXTRACTION BY ON-LINE SIGNAL SEPARATION AND RECOVERY" IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: ANALOG AND DIGITAL SIGNAL PROCESSING, IEEE INC. NEW YORK, US, vol. 46, no. 7, July 1999 (1999-07), pages 915-922, XP000919887 ISSN: 1057-7130 abstract; figure 1	1-35

INTERNATIONAL SEARCH REPORT

Information on patent family members

Int tional Application No
PCT/US 01/31121

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5630015 A	13-05-1997	DE 69131739 D1	02-12-1999
		DE 69131739 T2	04-10-2001
		DE 69132644 D1	26-07-2001
		DE 69132645 D1	26-07-2001
		DE 69132659 D1	16-08-2001
		DE 69132749 D1	31-10-2001
		EP 0459382 A2	04-12-1991
		EP 0763810 A1	19-03-1997
		EP 0763811 A1	19-03-1997
		EP 0763812 A1	19-03-1997
		EP 0763813 A1	19-03-1997
		JP 4230796 A	19-08-1992
		KR 9513552 B1	08-11-1995
		US 5621850 A	15-04-1997
		US 5617505 A	01-04-1997
		US 5355431 A	11-10-1994
		KR 9607843 B1	12-06-1996
		KR 9501070 B1	08-02-1995
		KR 9501071 B1	08-02-1995
		KR 9501067 B1	08-02-1995